Comparison of Different Methods of Outlier Detection in Univariate Time Series Data

Egbo Mary Nkechinyere
E-mail Address: egbomary4@yahoo.com
Department of Statistics, Federal University of Technology Owerri Nigeria
Owerri Nigeria

Iheagwara Andrew I.
E-mail Address: andyiheagwar@yahoo.com
Procurement Officer/Director Planning, Research & Statistics, Nigeria Erosion & Watershed Management Project (World Bank-Assisted), Ministry of Petroleum & Environment, Ploy 36, chief Executive Quarters, Area "B", New Owerri, Imo State Nigeria

Okenwe Idochi
E-mail Address: nwond@yahoo.com
Department of Statistics, School of Applied Sciences, Rivers State Polytechnic
PMB 20, Bori, Rivers State Nigeria

ABSTRACT

Overtime, different methods of detecting outliers have been worked on, some detected single outliers while others detected multiple outliers, some detected outliers in univariate models while others are limited to multivariate models, some others used simple measures while a lot others used the robust measures for detecting outliers. With these numerous methods raised the problem of which method is the best given a particular set of data. The best methods are subjective to the kind of data that is under consideration in the given study. For this study, we confined our attention to univariate time series data, subjected it to different methods of outlier detection in univariate data, detected the outliers and then worked on the efficiency of these different methods of outlier detection. We as well took time to outline the procedures of detecting univariate outlier in some common statistical software packages. It can be concluded from the evidence of this study that the 3SD method and the Z-score method of outlier detection is not a good model for detecting outliers in univariate model. This can be attributed to the parameters they use for estimation of outliers in these data sets.

Key words: Univariate time series data, outlier detection, MADe Rule, Modified Z-Score, 2SD method, 3SD method

1. Background of the Study

No observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. The probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions, an observation that appears to stand apart from the bulk of other observations is called an "outlier", "extreme observations", "discordant observations", "rogue values". Ranjit(2001).

From time immemorial, outliers have always existed in all spheres of human endeavor. We only need to look at the ripe old ages of the people of old and that which is obtainable these days to reach these conclusion; in the old testament people lived very long life and one who dies at 80yrs of age is said to have died young

while one who dies at 500yrs of age died at a ripe old age. Nowadays, 80yrs is considered a ripe old age while nobody lives up to 500yrs any longer. With respect to both periods, dying at 80yrs or living up to 500yrs is considered an outlier when ripe old age is the factor under consideration (Human Longevity Facts).

Measuring the height of students in a particular class, we might find out that the heights of majority are evenly normally distributed besides a very few who might either be too tall or too short. This very group of people might adversely affect the mean height of the given population and can be classified as outliers.

Assume that the demand of "hand sanitizers'' is being surveyed in Nigeria for the previous three years, it will be noticed that demand for hand sanitizers will be relatively stable bar the period of march-October of 2014 where there is a fire sale of the product. This fire sale can be attributed to incidence of the Ebola viral disease in the country. Thus the demand for the hand sanitizers most likely would have skyrocketed to an alarming level and may really affect the measure gotten from other periods. This is another practical example of an outlier.

Given a mean and a standard deviation, a statistical distribution expects value to fall within a certain range. Discordant values that fail to conform to these ranges are called outliers. An outlier is a value that appears to deviate markedly from other member values of the sample in which it occurs or it can be defined as a point which is significantly different from others.

Error in data is one of the facts that cause parameter estimations to be subjective. If the erroneous case is proved statistically, then these cases are called outliers. Outliers are defined as the few observations or records in a data which appears to be inconsistent with the rest of the group of the sample and more effective on prediction values.

Hawkins(1980) articulately defined an outlier thus "an outlier is an observation which deviates so much from the observation as to arouse suspicions that it was generated by another mechanism". By this definition he tries to explain that normal data objects follow a "generating mechanism". Some statistical process and the abnormal objects deviate from this generating mechanism.

This is supported by a lot of other definitions. Barnett and Lewis (1994) described an outlier as "an observation which appears to be inconsistent with the remainder of that set of data''. Hampel (2001) evaluated concept of outliers without clear boundaries and some methods like the Grubbs rule can barely detect one outlier out of ten. Jagadish et al (1999) also stated each of them gives a different answer and none of them is conceptually satisfying. Tietjen-Moore (1972) generalized the Grubbs test to a case of detecting multiple outliers in the univariate data sets that are approximately normally distributed.

Areas of research such as statistics, data mining, information theory and process control theory have produced various methods for spotting outliers in stochastic processes. Some of these methods include the turkey's methods, the standard deviation method, Z-score method, the modify Z-score method, the MAD_E method, and the median rule to mention but a few. While graphically, outlying observations can be spotted using the normal probability plot, the run sequence plots, histogram, box plots, or lag plots.

It is important as well to differentiate between the parametric and non parametric procedures. Parametric procedures assumes the value to be identically and independently distributed following a known probability distribution (generally a normal distribution) while the non parametric procedures are the model free procedures which are often unsuitable or generally imprecise for data sets without prior knowledge of the underlying distribution because the hypothesis (e.g independence of values) are not satisfied and because the statistical models are not reliable for real data sets and are hard to validate since many data sets do not fit one particular model. The non-parametric methods do not assume knowledge of statistical distributions.

Although outliers are considered an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise; adversely lead to model specification, biased parameter estimation and incorrect results. It is therefore paramount to identify them before modeling and analysis (Williams et al, 2002), (Liu et al, 2004). Several outlier detection methods have been developed. Some methods are sensitive to extreme values, like the SD method and Z-score method, and others are resistant to extreme values, like turkey's method. Although these methods are quite powerful with large normal data, it may be problematic to apply them to non normal data or small sample sizes without knowledge of their characteristics in these circumstances. This is because each labeling method has different measures to detect outlier, and expected outlier percentages change differently according to the sample size or distribution type of the data. Hence, this study shall examine to know the best method of detecting outliers in a univariate time series data using eight different methods namely; the 2 Standard Deviation method, 3 standard deviation method, the Z-score method, the Modified Z-score method, the Tukey's method, the median rule, the 2 Median Absolute Deviation Method, and 3 Median Absolute Deviation Method.

1.1. Statement of the Problem

Outliers have been a major problem in the area of statistics, including modeling, analysis and forecasting. Lots of methods has been portrayed as a means of detecting outliers in univariate time series data but not many works has been done on which of this methods is best for detecting outliers in univariate models. This work addresses that problem by using eight univariate outlier detection methods and checking which of them is best or more efficient in detecting outliers.

1.2. Related Literature Review

Ahmet (2010) carried out a work on statistical modelling for outlier factors. In his study, he was concerned with outliers in time series which have two special cases, innovational outlier (IO) and additive outlier (AO). The occurrence of AO indicates that action is required, possibly to adjust the measuring instrument or mistake made by person in observation or record. However, if IO occurs, no adjustment of the measurement operation is required. Also in the study, a multi-factor (3^2 42**)** modelling was done in order to fit the effects of model in data analysis AR(1) coefficients, (0.5, 0.7, 0.9) outlier type (AO, IO), series wideness (50, 100, 200, 500) and criterion value sensibility (% 99 (C=3.00), % 95 (C=3.50), % 90 (C=4.00)) factors statistically by making use of a simulation study. The results of the variance analysis on outlier factors were also emphasized.

Regina and Agustin (2001) worked on Seasonal outliers in time series. The standard procedures for automatic outlier detection and correction considered four types of outliers, namely, the additive, innovational, level shift, and transitory change outliers. In their study, it was argued that typification presented serious shortcomings. First, the innovational outlier may display undesirable features; second, it was incomplete because it couldn't model breaks in the pattern of the seasonal component. Several specifications for a seasonal outlier were considered and the one denoted Seasonal Level Shift (SLS) was analyzed in detail through simulation and real examples. It was concluded that the SLS displays better properties and turns out to be more useful than the innovational outlier, and hence the typification of outliers in automatic outlier detection and correction should replace the latter type of outlier by the seasonal level shift one.

Hau and Tong (1989), gave a practical method for outlier detection in autoregressive models. They achieved this by using a mahalanobis distance function which requires minimal computation after the data has been fitted. According the researchers, the practical method can be used to detect both innovative and additive outliers and applies to both real and simulated data.

Yamanishi and Takeuchi (2002), was concerned with detecting outliers and change point detection from data stream. They proposed a unifying framework as criteria for dealing with outliers and change point problems. In this frame work, a probabilistic model of the data source is incrementally learned using an on-line discounting learning algorithm which can track the changing data source adaptively by forgetting the effect of past data gradually. Then the score for any given data were calculated to measure its deviation from the learned model with a higher score indicating a higher possibility of being an outlier.

Ferdousi and Maeda (2006) tackled the problem of finding outliers in time series financial data using Peer Group Analysis (PGA). The PGA is an unsupervised fraud detection technique whose main objective is to characterize the expected pattern of behavior around the target sequence in terms of behavior of similar object, and then to detect any difference in evolution between the expected pattern and the target. The technique was applied to stock market data and t-statistic was used to find the deviations effectively.

Chawla and Sun (2006), delivered a work on outlier detection as a core data mining paradigm. In their work, they listed different methods of outlier detection, explained outlier detection as an unsupervised learning and gave a classical and modern statistical approaches in detecting outliers.

Olewuezi (2011), compared three well known outlier labeling methods namely; standard deviation method, median absolute deviation method, and median rule as a guideline for determining the best choice of outlier detection. She concluded from the result of the estimated outliers that the standard deviation method is inappropriate to use here because it is highly sensitive to extreme values.

Gupta et al (2014), enlisted the various forms of temporal data collection, how to manage these temporal data and then provided a comprehensive and structured definition of outliers in temporal data, as well as how to detect these outliers.

In this study, we focused on the comparison of the methods of detecting outliers in a univariate time series data. We used eight methods of outlier detection namely; the 2 standard deviation method, 3 standard deviation method, the Z-score method, the modified Z-score method, the Boxplot method, the median rule, the 2 Median Absolute Deviation Method and the 3 Median Absolute deviation method. The data employed is the inflation rate in Nigeria for a thirty years period. We as well tried to detect which of these methods is more efficient for outlier detection.

2. Methodology

Data Sources

The data is a data on inflation rate in Nigeria over a thirty three years period ranging from 1981 to 2013. The data was gotten from the Central Bank of Nigeria yearly statistical bulletin which was published on April 2014.

Methods of Analysis

Standard Deviation Method of Outlier Detection

Here we consider outlier detection using the standard deviation method. The basic idea behind it is that Y_i‘s; i = 1,2,……,n, follows a normal distribution, then this is a simple classical approach to screen outliers. It can be 2 Standard Deviation (2SD) or 3 Standard Deviation (3SD) depending on the researcher.

It is defined as:

2SD method: $\bar{Y} \pm 2SD$ (1)

3SD method: $\bar{Y} \pm 3SD$ (2)

Where $\bar{Y}$ is the sample mean and SD the sample standard deviation

The observations outside these intervals are considered as outliers. To determine the number of data that are likely to be out of these ranges, we employ the Chebyshev inequality for a random variable Y with mean μ and variance δ^2, which states:

$P[(Y - \mu) < K\delta] \geq 1 - 1/K^2; k>0$ (3a)

$P[\ /Y - \mu/ \geq K\delta\ ] \leq 1/K^2$ (3b)

The inequality allows us to know the proportion of our data that will be within K-standard deviation of the mean.

For a given random variable Yi which takes different values at different points $Y_1, Y_2, \ldots\ldots, Y_n$:
The mean = $\bar{Y}$; and the standard deviation is estimated thus;

Table 1 $(Y_1 - \bar{Y})^2$

$\mathbf{Y_i}$	$\mathbf{Y_i - \bar{Y}}$	$\mathbf{(Y_i - \bar{Y})^2}$
Y_1	$Y_1 - \bar{Y}$	$(Y_1 - \bar{Y})^2$
Y_2	$Y_2 - \bar{Y}$	$(Y_2 - \bar{Y})^2$
.	.	.
.	.	.
Y_n	$Y_n - \bar{Y}$	$(Y_n - \bar{Y})^2$

$$Variance\left(\sigma^2\right) = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{y}\right)^2}{n-1} \quad \text{Std. Dev} = \sqrt{\sigma^2}$$

For 2SD Method: $\bar{Y} \pm 2SD = [I_1, I_2]$

For 3SD method: $\bar{Y} \pm 3SD = [I_3, I_4]$

Where $I_1$and I_2 represent the various intervals. Values of Y that fall outside the intervals are considered outliers.

The Z-score Method of Outlier Detection

Another popular method of detecting outliers is the Z-score technique which makes use of the variable values, the mean and the standard deviation. Just like in the SD method, it still assumes that the Yi's follows a normal distribution and the Z follows a standard normal distribution N(0,1). Z-score is computed as:

$Z_i = (Y_i - \bar{Y})/\delta$ (4)

The rationale behind this is that any Z-score whose absolute value is ≥ 3, is considered an outlier while any observation whose value is relatively less than 3 are not outliers.

The problem with these method is that in the presence of more than one extreme value, a masking problem occurs in which outlying points are considered non-outliers in the presence of other outliers.Following from table 1, Z-score is estimated as thus;

Table 2

$(Y_i - \bar{Y})$	$Z_i=(Y_i - \bar{Y})/\delta$	$/Z_i/$
$(Y_1 - \bar{Y})$	$(Y_1 - \bar{Y})/\delta$	$/Z_1/$
$(Y_2 - \bar{Y})$	$(Y_2 - \bar{Y})/\delta$	$/Z_2/$
. . . $(Y_n - \bar{Y})$	. . . $(Y_n - \bar{Y})$	. . . $/Z_n/$

For any $/Z_i/ \geq 3$, Y_i is an outlier. (5)

The Modified Z-Score Method of Outlier Detection

Just like in the previous two methods already studied, this method assumes the data follows a normal distribution but unlike the previous two, it does not make use of the sample mean and standard deviation since those two measures are largely affected by even a single extreme value. But it makes use of the median and the median absolute deviation (MAD).

MAD = median $[/Y_i - \tilde{Y}/]$ (6)

where $\tilde{Y}$ is the data series median.

According to Iglewicz and Hoaglin (1993), they suggested that the modified Z-score is computed by;

$$M_i = \frac{0.6745\left(Y_i - \tilde{Y}\right)}{MAD} \quad (7)$$

where $E(MAD) = 0.675\delta$ (8)

If $/Mi/ > 3.5$, (9)

Then Y_i is an outlier.

It is given as thus:

Table 3

I	$\mathbf{Y_i}$	$\mathbf{(Y_i - \tilde{Y})}$	$\mathbf{M_i = 0.6745\ (Y_i - \tilde{Y})/MAD}$	$\mathbf{/M_i/}$
1	Y_1	$(Y_1 - \tilde{Y})$	M_1	$/M_1/$
2	Y_2	$(Y_i - \tilde{Y})$	M_2	$/M_2/$
. . N	. . Yn	. . Yn - $\tilde{Y}$	. . Mn	. . /Mn/

The Box Plot Method of Outlier Detection

This is also known as the Tukey's method, proposed by Tukey (1977), is a very favorable method of detecting outliers because it makes no distributional assumption on the data.

It is a well known simple graphical tool used to display information about continuous univariate data, such as the median, lower quartile, lower extreme and upper extremes of a data sets.

The boxplot/Turkey's method uses the quartiles which are not affected by extreme values. The basics of the method are stated below;

I) The IQR is the distance between the lower quartile (Q_1) and upper quartile (Q_3)
II) The inner fences are located at a distance of 1.5IQR below Q_1 and above Q_3, [Q_1 – 1.5IQR, Q_3 + 1.5IQR] (10)
III) Outer fences are located at a distance of 3IQR below Q1 and above Q3 [Q_1 – 3IQR, Q_3 + 3IQR] (11)
IV) A value between the inner and outer fence is called a possible outlier
V) Values between the inner and outer fence are possibly outliers, data outside the outer fence are probably outliers

For a set of data Y_i; i = 1,2,3,…..,n

Let K denote the inter-quartile range, the inner fence is [Q_1 – 1.5k, Q_3 + 1.5], the outer fence is given an [Q_1 – 3k, Q_3 + 3k]. The values of Y_i not in the inner and outer fence are possible outliers while those not in the fence at all are extreme outliers even though this is mostly not obtainable.

Median Rule

If $Y_1,Y_2,…..,Y_n$ is a random sample of size n arranged in order of magnitude, then we define the define the median as

$\tilde{Y} = Y_m$; for n odd (12)

$[Y_m + Y_{m+1}]/2$; for n even (13)

Hence the median is the value that falls exactly in the centre of the data when the data are arranged in order. Carling (2000) introduced the identification of outliers through studying the relationship between target outlier percentages and generalized Lambda Distribution (GLD). GLD's with different parameters are used for various moderately skewed distributions.

The median rule introduced by carling (1998) is a robust estimator of location having approximately 50% breakdown point. The method is given by the range;

$\{I_1,I_2\} = Q_2 \pm 2.3$ IQR, (14)

where Q_2 = sample median, IQR = inter-quartile range.

The values of Y_i that fall outside the intervals $\{I_1, I_2\}$ is labeled an outlier.

The MADe Rule.

This is one of the basic robust methods of outlier detection, developed by Ratcliffe (1993), which is largely unaffected by the presence of extreme values in the data. This approach is similar to the standard deviation method. However, the median and the median absolute deviation are often employed in this method instead of the mean and standard deviation. The MADe method is defined as follows:

2MADe Method: Median ± 2MADe (15)

3MADe Method: Median ± 3MADe (16)

Where MADe = 1.483MAD for large normal data. (17)

This is because when it is scaled by a factor of 1.483, it is similar to the standard deviation method in a normal distribution.

Normality Test

The Anderson Darling test is used to test if a sample of a data comes from a population with a specific distribution. It is a modification of the Kolmogrov-Smirnov (K-S) test and gives more weight to the tails than the K-S test does. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson Darling test makes use of specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that the critical value must be calculated for each distribution.

After the data is plotted for normality test, we check the p-value;

If: p-value < 0.05, (normal)

p-value ≥ 0.05, (not normal)

Note: In the presence of outliers, the data is expected to be non-normal since outliers generally increase the error variance of the data thus making it not normal. But after the outliers have been detected and removed, the data is expected to become normally distributed.

In this work, normality was checked before detecting the outliers and as well after detecting and removing the outliers for each method.

Analysis of Data

TABLE 4. Nigerian's Inflation Rate (1981-2013)

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	1998	10.0
1982	7.7	1999	6.6
1983	23.2	2000	6.9
1984	39.6	2001	18.9
1985	5.5	2002	12.9
1986	5.4	2003	14.0
1987	10.2	2004	15.0
1988	38.3	2005	17.9
1989	40.9	2006	8.5
1990	7.5	2007	5.4
1991	13.0	2008	15.1
1992	44.5	2009	13.9

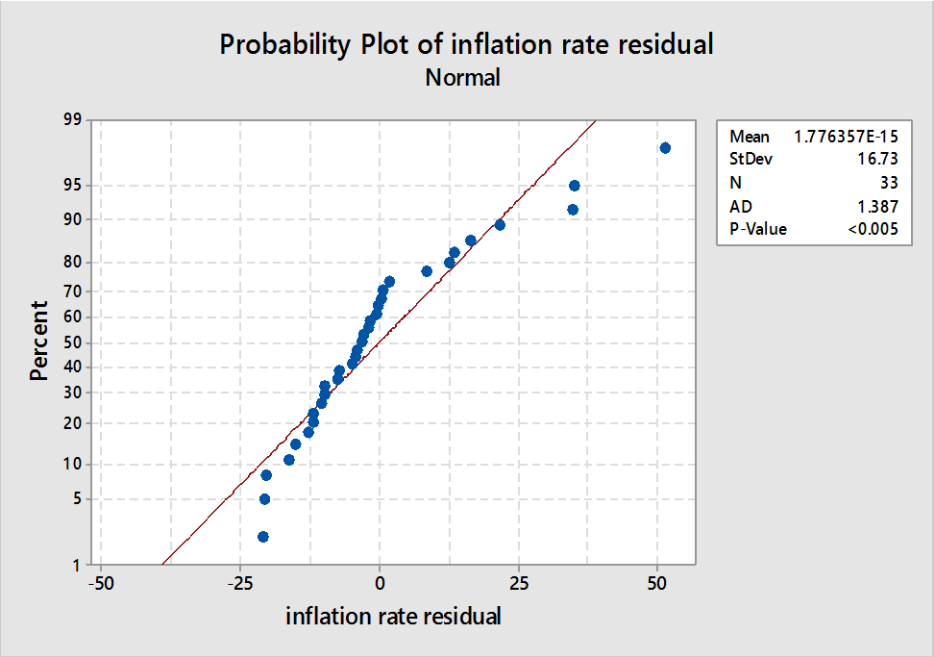
1993	57.2	2010	11.8
1994	57.0	2011	10.3
1995	72.8	2012	12.0
1996	29.3	2013	8.0
1997	8.5		

Source: Central Bank of Nigeria Annual Report, April 2014.

Normality Test

We tested the normality of the data using the Anderson Darling test of normality in minitab. This was done using the residual value of inflation rate. The following figure was obtained:

Figure 1: normality plot for inflation rate



The hypotheses are given thus:

H_0: the data is normally distributed

H_1: the data is not normally distributed

Testing at 0.025 level of significance

Conclusion: since P-value = 0.005 < α = 0.05, we reject H_0 and conclude that the data is not normal. This is the expected result because outliers are still present in the data.

Table 5: Computational Results Table.

I	Yi	Rank Yi	Yi - $\bar{Y}$	(Yi - $\bar{Y}$)2	/Z_i/= (/Yi - $\bar{Y}$/)/δ	/Yi - $\tilde{Y}$/	rank /Yi - $\tilde{Y}$/	/M_i/= 0.6745(/Yi - $\tilde{Y}$/)/MAD
1	20.9	24	0.640	0.410	0.037	7.90	24	0.969

2	7.7	7	-12.560	157.754	0.720	5.30	16	0.650
3	23.2	25	2.940	8.644	0.168	10.20	25	1.251
4	39.6	28	19.340	374.036	1.108	26.60	28	3.262
5	5.5	3	-14.760	217.858	0.846	7.50	21	0.920
6	5.4	1.5	-14.860	220.820	0.852	7.60	22.5	0.932
7	10.2	12	-10.060	101.204	0.577	2.80	10	0.343
8	38.3	27	18.040	325.442	1.034	25.30	27	3.103
9	40.9	29	20.640	426.010	1.183	27.90	29	3.422
10	7.5	6	-12.760	162.818	0.731	5.50	17	0.675
11	13.0	17	-7.260	52.708	0.416	0.00	1	0.000
12	44.5	30	24.240	587.578	1.389	31.50	30	3.863
13	57.2	32	36.940	1364.564	2.117	44.20	32	5.421
14	57.0	31	36.740	1349.828	2.105	44.00	31	5.396
15	72.8	33	52.540	2760.452	3.011	59.80	33	7.334
16	29.3	26	9.040	81.722	0.518	16.30	26	1.999
17	8.5	9.5	-11.760	138.298	0.674	4.50	12.5	0.552
18	10.0	11	-10.260	105.268	0.588	3.00	11	0.368
19	6.6	4	-13.660	186.596	0.783	6.40	20	0.785
20	6.9	5	-13.360	178.490	0.766	6.10	19	0.748
21	18.9	23	-1.360	1.850	0.078	5.90	18	0.724
22	12.9	16	-7.360	54.170	0.422	0.10	2	0.012
23	14.0	19	-6.260	39.188	0.359	1.00	4.5	0.123
24	15.0	20	-5.260	27.668	0.301	2.00	7	0.245
25	17.9	22	-2.360	5.570	0.135	4.90	14	0.601
26	8.5	9.5	-11.760	138.298	0.674	4.50	12.5	0.552
27	5.4	1.5	-14.860	220.820	0.852	7.60	22.5	0.932
28	15.1	21	-5.160	26.626	0.296	2.10	8	0.258
29	13.9	18	-6.360	40.450	0.364	0.90	3	0.110
30	11.8	14	-8.460	71.572	0.485	1.20	6	0.147
31	10.3	13	-9.960	99.202	0.571	2.70	9	0.331
32	12.0	15	-8.260	68.228	0.473	1.00	4.5	0.123
33	8.0	8	-12.260	150.308	0.703	5.00	15	0.613
TOTAL	**668.7**		**0.120**	**9744.437**	**25.334**	**381.30**		**46.761**

The Standard Deviation Method

Mean = 668.9/33 = 20.26

Variance = 9744.437/32 = 304.51

Standard Deviation = $\sqrt{304.51}$ = 17.45

The 2SD Method

2SD Method: $\bar{Y} \pm 2SD = (I_1, I_2)$ from Eqn. (1)

$= 20.26 \pm 2(17.45)$ $= 20.26 \pm 34.9 = (-14.64,55.16)$

Conclusion

From these result, the data points **57.2, 57,** and **72.8** corresponding to the years **1993, 1994,** and **1995** are regarded as outliers.

The 3SD Method

3SD Method: $\bar{Y} \pm 3SD$ $= (I_1,I_2)$ from Eqn.(2)

$= 20.26 \pm 3(17.45)$ $= 20.26 \pm 52.35$ $= (-32.09,72.61)$

Conclusion

From the above result, the data point **72.8** corresponding to the year **1995** is regarded an outlier.

The Z-Score Method

Z-score is defined as deviation from the mean divided by the standard deviation.

Mathematically Z is given as in the equation From Eqn.(4)

If $/Z_i/ \geq 3$; $Y_i$is considered an outlier. From Eqn.(5)

From TABLE 4, in the $Z_i$column, it is noticed that the data point **72.8** corresponding to the absolute Z-score value of **3.011** for the year **1995** is an outlier while.

The Modified Z-Score Method

The major difference between the Z-Score and the Modified Z-Score is that the modified version makes use of the median which is a robust estimator.

Mathematically it is given by:

$M_i = 0.6745(Yi - \tilde{Y}) / MAD.$ From Eqn.(7)

Where $\tilde{Y}$ = Sample Median

MAD = Median of $/Y_i - \tilde{Y}/$. From Eqn. (6)

If $/M_i/ > 3.5$; the observation is an outlier. From Eqn.(9)

From the table labeled Mi, the Mi values **3.86, 5.39, 5.42, 7.33** corresponding to years **1992,1993,1994,1995** and values **44.5, 57.2, 57.0, and 72.8** are considered outliers.

Box Plot Method

1st quartile = 8.25

3^{rd} quartile = 26.25

Inter-quartile range = 26.25 – 8.25 = 18.

Inner fence = [Q1 – 1.5IQR, Q3 + 1.5IQR]. From Eqn.(10)

= [8.25 – 1.5(18), 26.25 + 1.5(18)] = [-18.75, 53.25]

Outer fence = [Q1 – 3IQR, Q3 + 3IQR] FromEqn.(11)

= [8.25 – 3(18), 26.25 + 3(18)] = [-45.75, 80.25]

This implies that data values **57.2, 57.0 and 72.8** corresponding to the different years **1993, 1994, and 1995** are possible outliers from the data.

Median Rule

The median $\tilde{Y}$ is denoted as the middle value. From the data,

$\tilde{Y} = 13$; IQR = 18

From the median rule,

$[I_1, I_2]$ is as given in the equation From Eqn. (14)

= 13 ± 2.3(18) = [-28.4,54.4]

Using the median rule, the values **57.2, 57, and 72.8** corresponding to the years **1993, 1994, and 1995** are considered outliers.

Median Absolute Deviation (MADe) Method

MAD is as given in the equation fromEqn.(16)

MAD = 5.5

MADe = 1.483MAD = 1.483(5.5) = 8.157 fromEqn.(17)

The 2MADe Method

Median ± 2MADe fromeqn. (15)

13 ± 2(8.157) = [-3.31, 29.31]

The data points **39.6, 38.3, 40.9, 44.5, 57.2, 57.0, 72.8** corresponding to the years **1984, 1988, 1989, 1992, 1993, 1994,1995** are considered outliers.

The 3MADe METHOD

Median ± 3MADe from eqn. (16)

13 ± 3(8.157) = [-11.47,37.47]

The data points **39.6, 38.3, 40.9, 44.5, 57.2, 57.0, 72.8** corresponding to the years **1984, 1988, 1989, 1992, 1993, 1994,1995** are considered outliers.

In the next page, we give a summary table of the outliers detected by this method so that a visual comparison can be made among all these methods of outlier detection.

Table 6: Outliers Summary Table.

Year	Inflation rate (%)	2 SD METHOD	3 SD METHOD	Z-SCORE METHOD	MODIFIED Z-SCORE	BOX PLOT METHOD	MEDIAN RULE	2 MADe METHOD	3MADe METHOD
1981	20.9	20.9	20.9	20.9	20.9	20.9	20.9	20.9	20.9
1982	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7	7.7
1983	23.2	23.2	23.2	23.2	23.2	23.2	23.2	23.2	23.2
1984	39.6	39.6	39.6	39.6	39.6	39.6	39.6	**39.6****	**39.6****
1985	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
1986	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4
1987	10.2	10.2	10.2	10.2	10.2	10.2	10.2	10.2	10.2
1988	38.3	38.3	38.3	38.3	38.3	38.3	38.3	**38.3****	**38.3****
1989	40.9	40.9	40.9	40.9	40.9	40.9	40.9	**40.9****	**40.9****
1990	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5
1991	13.0	13.0	13.0	13.0	13.0	13.0	13.0	13.0	13.0
1992	44.5	44.5	44.5	44.5	**44.5****	44.5	44.5	**44.5****	**44.5****
1993	57.2	**57.2 ****	57.2	57.2	**57.2****	**57.2****	**57.2****	**57.2****	**57.2****
1994	57.0	**57****	57.0	57.0	**57****	**57****	**57****	**57****	**57****
1995	72.8	**72.8****	**72.8****	**72.8****	**72.8****	**72.8****	**72.8****	**72.8****	**72.8****
1996	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3	29.3
1997	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5
1998	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
1999	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
2000	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9
2001	18.9	18.9	18.9	18.9	18.9	18.9	18.9	18.9	18.9
2002	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9
2003	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0
2004	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0
2005	17.9	17.9	17.9	17.9	17.9	17.9	17.9	17.9	17.9
2006	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5
2007	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4	5.4
2008	15.1	15.1	15.1	15.1	15.1	15.1	15.1	15.1	15.1
2009	13.9	13.9	13.9	13.9	13.9	13.9	13.9	13.9	13.9
2010	11.8	11.8	11.8	11.8	11.8	11.8	11.8	11.8	11.8
2011	10.3	10.3	10.3	10.3	10.3	10.3	10.3	10.3	10.3
2012	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0
2013	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0

Note: boxes with bold letters and double stars are detected outliers by the various methods

Outlier Results from Some Statistical Software Packages Using Given Data

Table 7: Table of Descriptive From SPSS

Descriptive

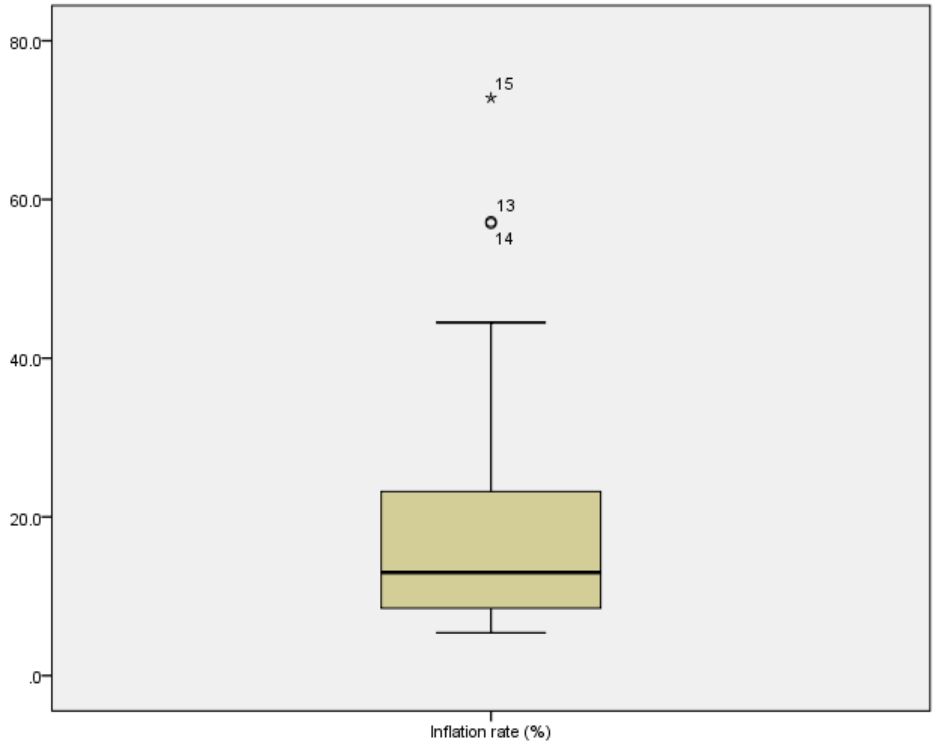
			Statistic	Std. Error
Inflation rate (%)	Mean		20.264	3.0377
	95% Confidence Interval for Mean	Lower Bound	14.076	
		Upper Bound	26.451	
	5% Trimmed Mean		18.512	
	Median		13.000	
	Variance		304.514	
	Std. Deviation		17.4503	
	Minimum		5.4	
	Maximum		72.8	
	Range		67.4	
	Interquartile Range		18.0	
	Skewness		1.578	.409
	Kurtosis		1.761	.798

Table 8: Table of Extreme Values from SPSS

Extreme Values

			Case Number	Value
Inflation rate (%)	Highest	1	15	72.8
		2	13	57.2
		3	14	57.0
		4	12	44.5
		5	9	40.9
	Lowest	1	27	5.4
		2	6	5.4
		3	5	5.5
		4	19	6.6
		5	20	6.9

Figure 2: SPSS box plot for inflation rate indicating outlying points

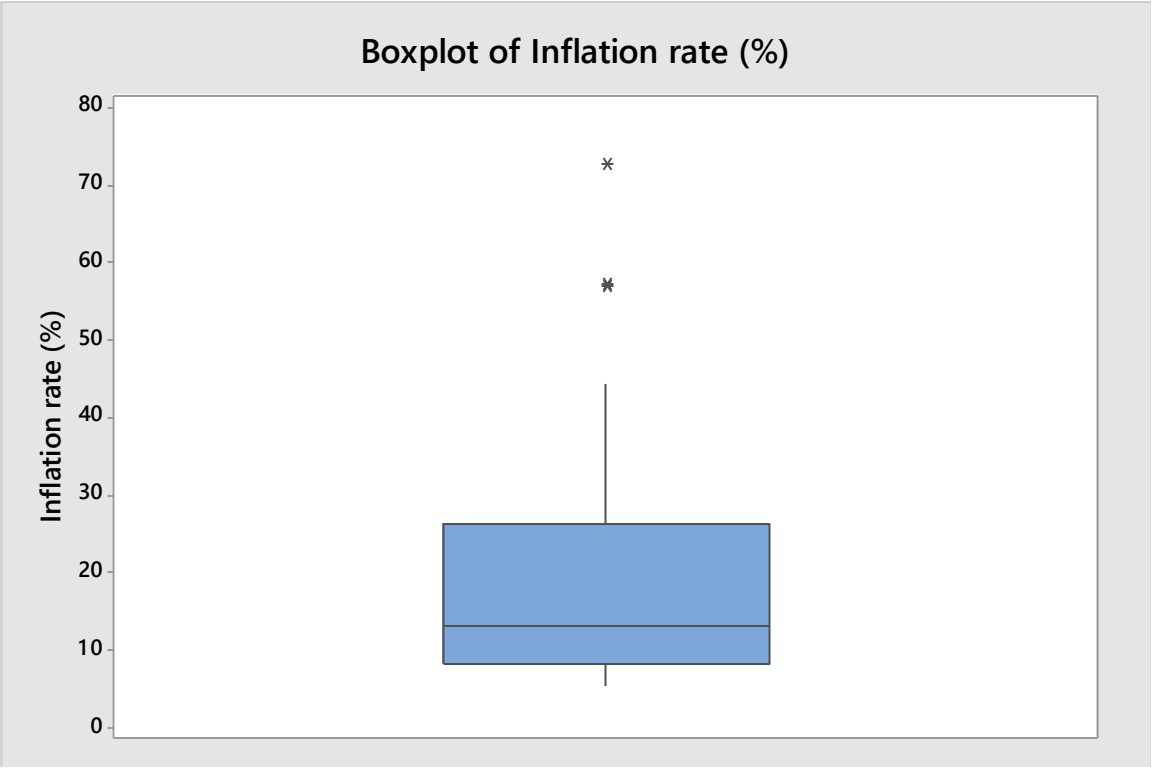


The first table gives us a table of descriptive statistics contains the mean, standard deviation, range, median, inter-quartile range etc. The next table is a table of extremes that brings out the five extreme values for us from both the maximums and minimums.

The next is the boxplot proper that indicates outliers by marking points outside the lower and upper limits of the boxplot. From here we can see that data entry 13, 14, and 15 corresponding to the inflation rate value of 57.2, 57, and 72.8 of the years 1993, 1994, 1995 were detected as outliers.

Results from Minitab

Figure 3: Minitab box plot for inflation rate indicating outlying points



This gives a simple boxplot with the interquartile range box and the outlier symbols. From the results we can notice that data points 57.2, 57, and 72.8 corresponding to the years 1993, 1994, and 1995 are detected as outliers using the box plot in minitab.

Efficiency of Outlier Detection Methods Considered

Efficiency of the Standard Deviation Method

The table is drawn again when the outlier detected by the standard deviation method has been removed.

From the 2SD Method:

Table 9

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	1999	6.6
1982	7.7	2000	6.9
1983	23.2	2001	18.9
1984	39.6	2002	12.9
1985	5.5	2003	14.0
1986	5.4	2004	15.0
1987	10.2	2005	17.9
1988	38.3	2006	8.5
1989	40.9	2007	5.4

1990	7.5	2008	15.1
1991	13.0	2009	13.9
1992	44.5	2010	11.8
1996	29.3	2011	10.3
1997	8.5	2012	12.0
1998	10.0	2013	8.0

The descriptive table for these is given below:

Table 10

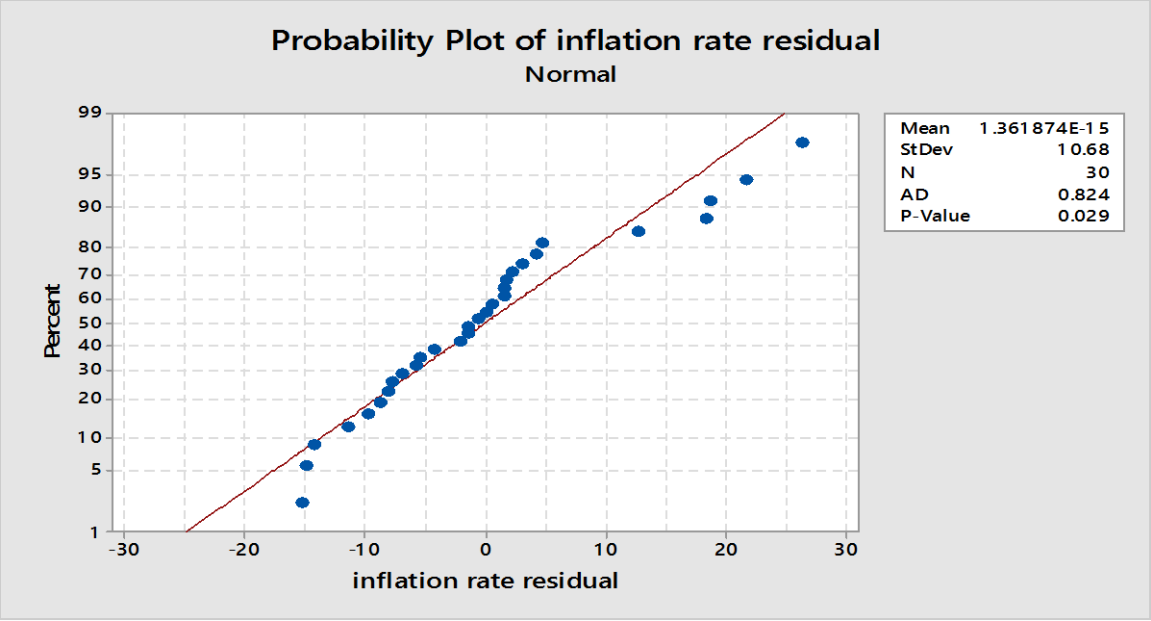
SDescriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	30	5.4	44.5	16.057	2.0732	11.3556
Valid N (listwise)	30					

Normality Test

This test if the data has normalized after the outlier detected by the 2SD method has been removed. The table is given below

Figure 4: normal plot for the 2SD method of outlier detection



The hypothesis and significance level retained, we reject H_0 and conclude that the data is approximately not normally distributed at 0.05 level of significance. Since P-value = 0.029 < α =0.05

From the 3SD method:

The table for inflation rate is given again when the outliers from the 3SD method has been detected and removed.

Table 11

Year	Inflation rate (%)	Year	Inflation rate
1981	20.9	1998	10.0
1982	7.7	1999	6.6
1983	23.2	2000	6.9
1984	39.6	2001	18.9
1985	5.5	2002	12.9
1986	5.4	2003	14.0
1987	10.2	2004	15.0
1988	38.3	2005	17.9
1989	40.9	2006	8.5
1990	7.5	2007	5.4
1991	13.0	2008	15.1
1992	44.5	2009	13.9
1993	57.2	2010	11.8
1994	57.0	2011	10.3
1996	29.3	2012	12.0
1997	8.5	2013	8.0

The descriptive table for these is given below

Table 12

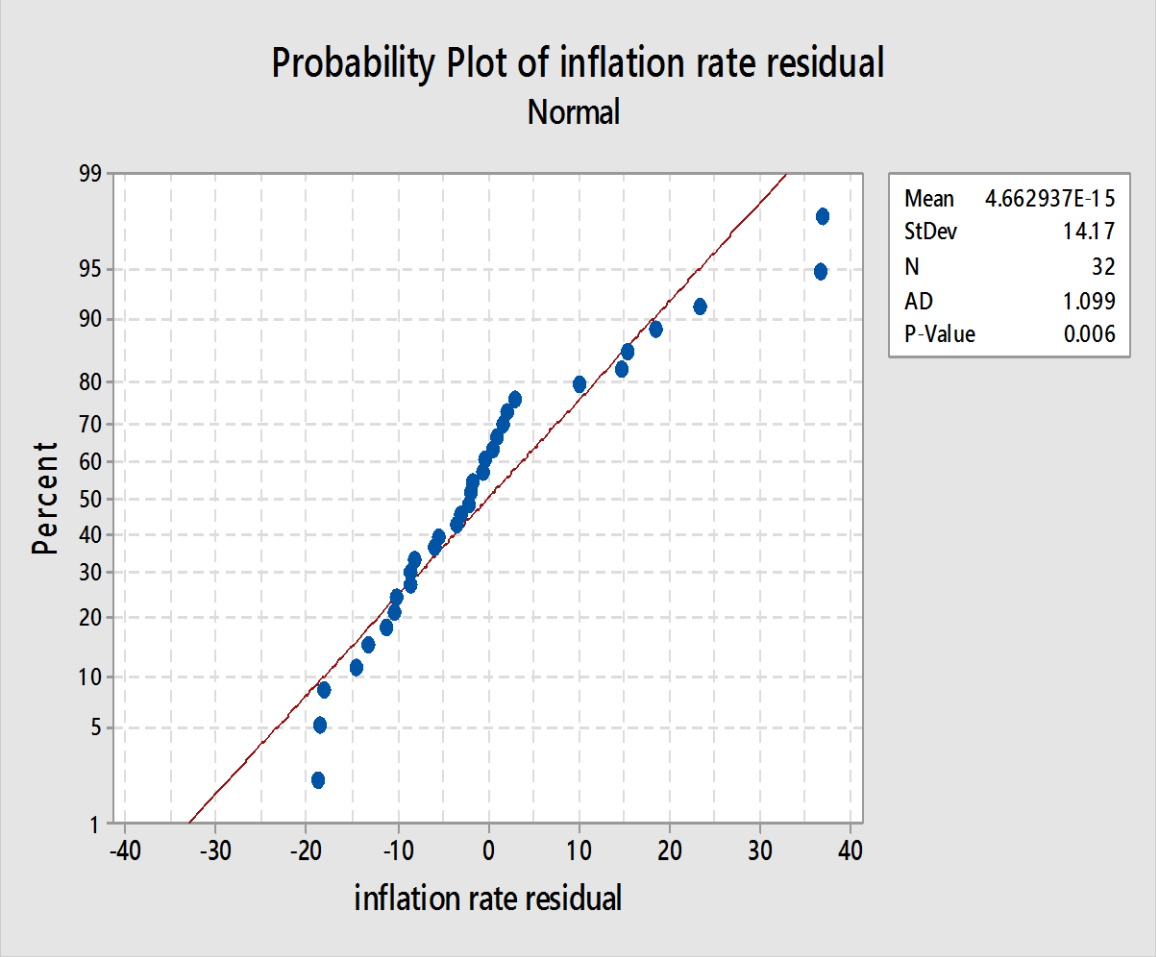
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	32	5.4	57.2	18.622	2.6370	14.9171
Valid N (listwise)	32					

Normality Test

This tests if the data has become normal after the outliers detected by the 3SD method have been removed.

Figure 5: normal plot for 3SD method of outlier detection



The hypothesis and significance level retained, we reject H_0 and conclude that the data is non-normal even after removing the outliers detected by the 3SD method. Since P-value = 0.006 < α = 0.05

Efficiency of the Z-Score Method of Outlier Detection

The data is given again when the outlier from the Z-Score method has been removed from the data set.

Table 13

Year	Inflation rate (%)	year	Inflation rate
1981	20.9	1998	10.0
1982	7.7	1999	6.6
1983	23.2	2000	6.9
1984	39.6	2001	18.9
1985	5.5	2002	12.9
1986	5.4	2003	14.0
1987	10.2	2004	15.0
1988	38.3	2005	17.9
1989	40.9	2006	8.5
1990	7.5	2007	5.4
1991	13.0	2008	15.1

1992	44.5	2009	13.9
1993	57.2	2010	11.8
1994	57.0	2011	10.3
1996	29.3	2012	12.0
1997	8.5	2013	8.0

The descriptive statistics for the data is given below:

Table 14

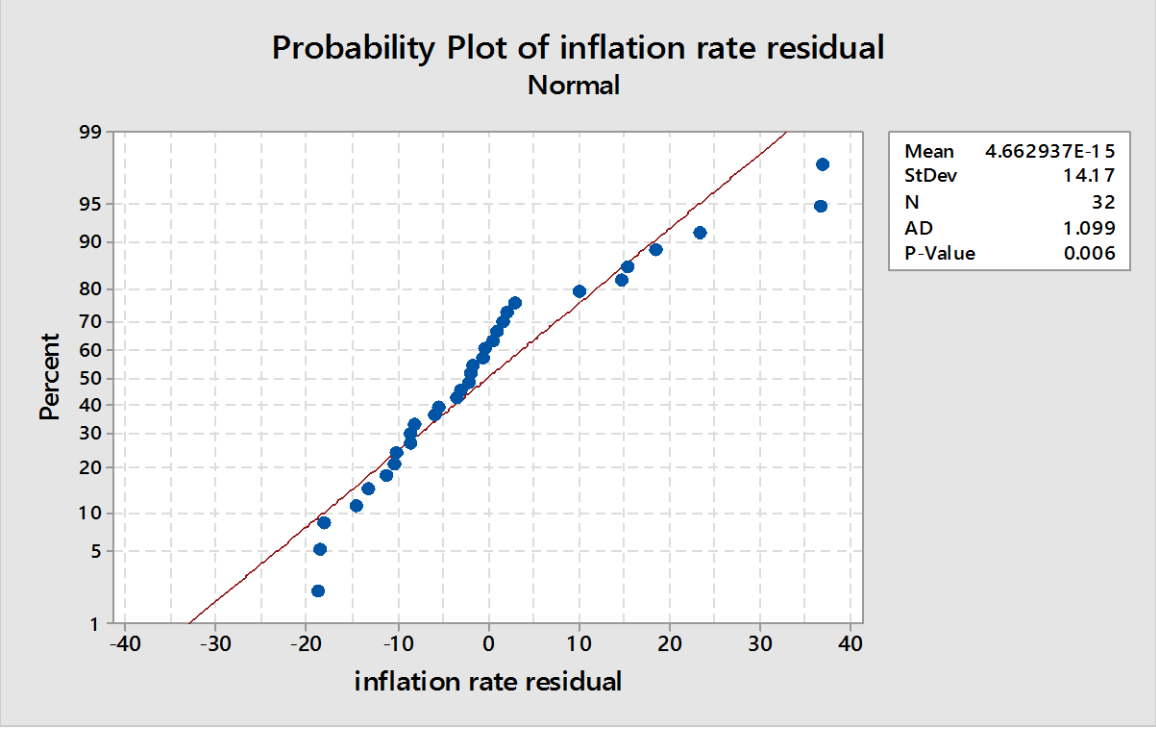
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	32	5.4	57.2	18.622	2.6370	14.9171
Valid N (listwise)	32					

Normality Test

This tests if the data has become normal after the outliers detected by the Z-Score method has been removed.

Figure 6: normal plot for outlier detection using Z-score method



The hypothesis and significance level retained, we reject H_0 and conclude that the data is non-normal even after removing the outliers detected by the 3SD method. Since P-value = 0.006 < α = 0.05

Efficiency of the Modified Z-Score Method of Outlier Detection

The data for inflation rate is given below after the outliers by the method of modified Z-Score have been removed.

Table 15

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	2000	6.9
1982	7.7	2001	18.9
1983	23.2	2002	12.9
1984	39.6	2003	14.0
1985	5.5	2004	15.0
1986	5.4	2005	17.9
1987	10.2	2006	8.5
1988	38.3	2007	5.4
1989	40.9	2008	15.1
1990	7.5	2009	13.9
1991	13.0	2010	11.8
1996	29.3	2011	10.3
1997	8.5	2012	12.0
1998	10.0	2013	8.0
1999	6.6		

The descriptive statistics of the above table is given as:

Table 16

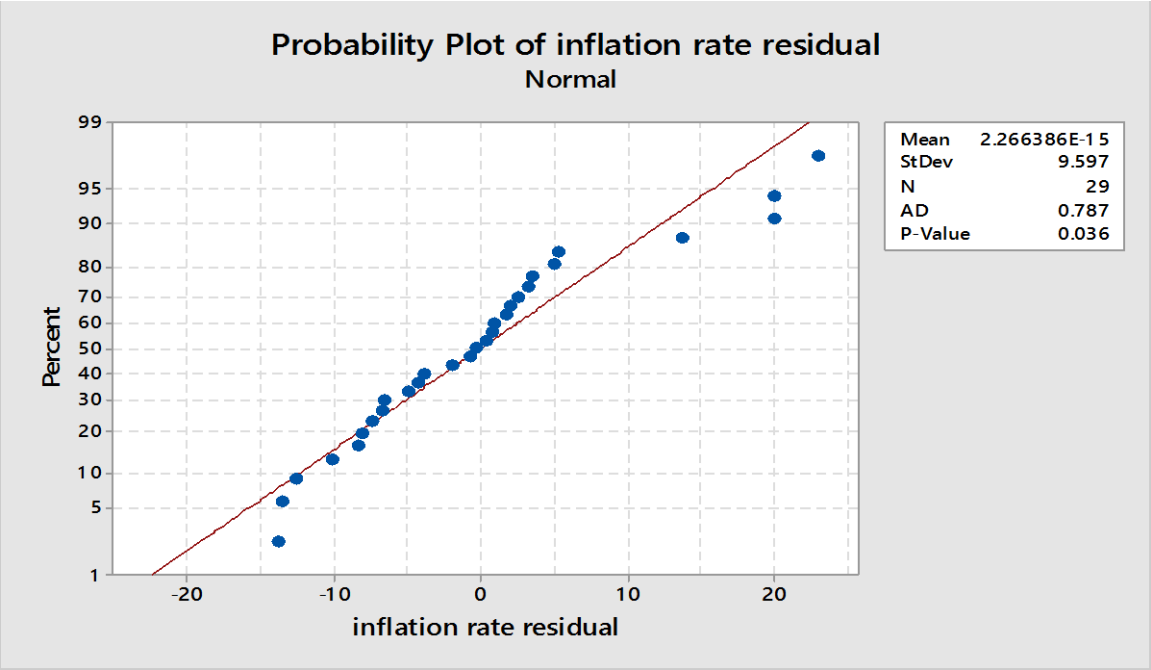
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	29	5.4	40.9	15.076	1.8907	10.1816
Valid N (listwise)	29					

Normality Test

The data is tested for normality after the outliers have been detected using the modified Z-score method of outlier detection. The figure is given below:

Figure 7: normal plot for modified Z-score method of outlier detection



The hypothesis and significance level retained, we reject H_0 and conclude that the data is approximately not normally distributed at 0.05 level of significance. Since P-value = 0.036 < α =0.05

Efficiency of the Box Plot Method of Outlier Detection

The table of inflation rate is given after the outliers detected by the box plot method have been removed.

Table 17

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	1999	6.6
1982	7.7	2000	6.9
1983	23.2	2001	18.9
1984	39.6	2002	12.9
1985	5.5	2003	14.0
1986	5.4	2004	15.0
1987	10.2	2005	17.9
1988	38.3	2006	8.5
1989	40.9	2007	5.4
1990	7.5	2008	15.1
1991	13.0	2009	13.9
1992	44.5	2010	11.8
1996	29.3	2011	10.3
1997	8.5	2012	12.0
1998	10.0	2013	8.0

The descriptive table for these is given below:

Table 18

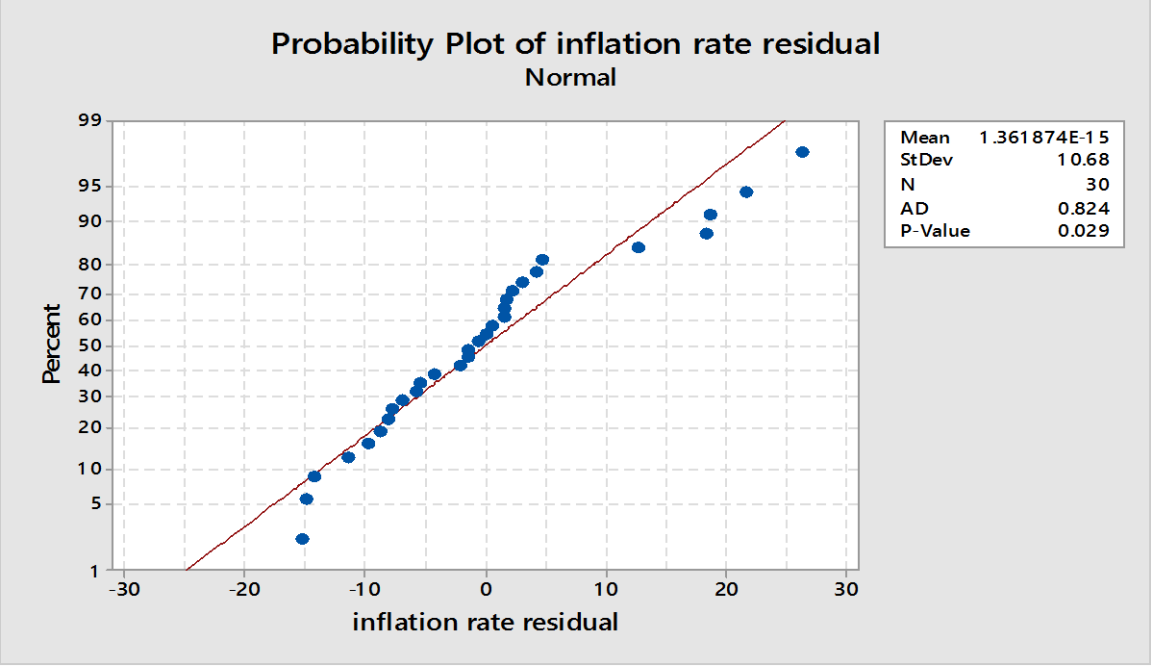
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	30	5.4	44.5	16.057	2.0732	11.3556
Valid N (listwise)	30					

Normality Test

These test if the data has normalized after the outliers detected by the boxplot method have been removed. The table is given below:

Figure 8: normal plot for box plot method of outlier detection



The hypothesis and significance level retained, we reject H_0 and conclude that the data is approximately not normally distributed at 0.05 level of significance. Since P-value = 0.029 < α =0.05

Efficiency of the Median Rule Method of Outlier Detection

The table of inflation rate is given below when the outliers detected by the median rule method has been removed:

Table 19

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	1999	6.6
1982	7.7	2000	6.9
1983	23.2	2001	18.9
1984	39.6	2002	12.9
1985	5.5	2003	14.0
1986	5.4	2004	15.0
1987	10.2	2005	17.9
1988	38.3	2006	8.5
1989	40.9	2007	5.4
1990	7.5	2008	15.1
1991	13.0	2009	13.9
1992	44.5	2010	11.8
1996	29.3	2011	10.3
1997	8.5	2012	12.0
1998	10.0	2013	8.0

The descriptive table for these is given below:

Table 20

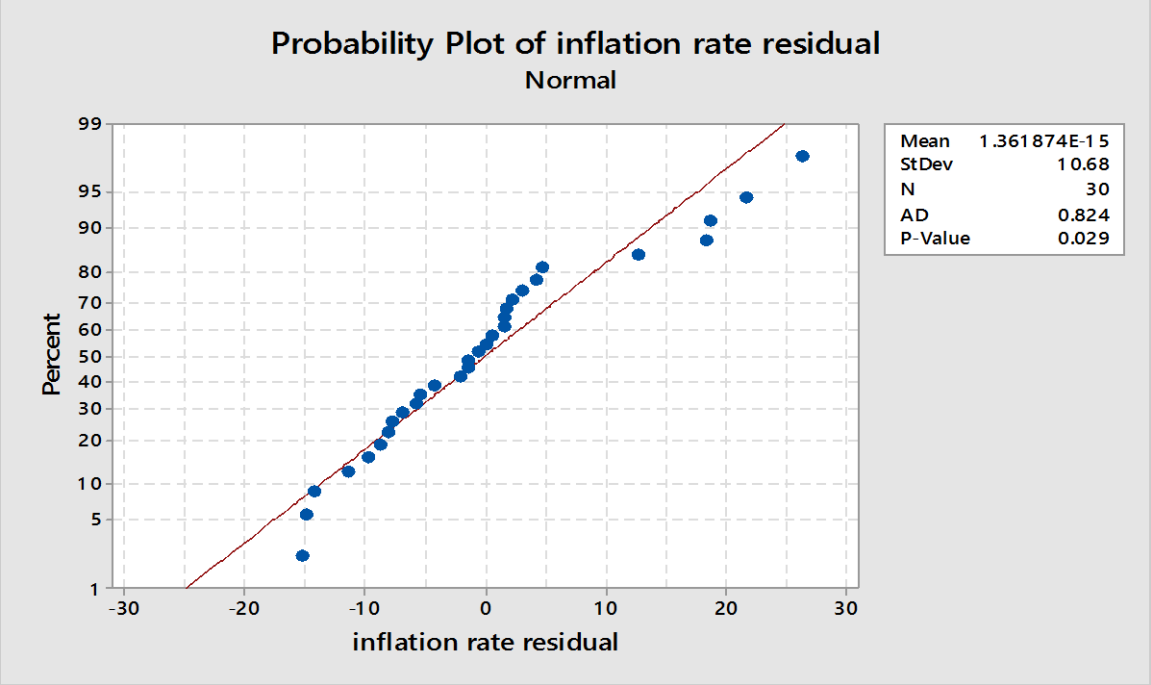
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	30	5.4	44.5	16.057	2.0732	11.3556
Valid N (listwise)	30					

Normality Test

This test if the data has normalized after the outlier detected by the median rule method has been removed. The table is given below:

Figure 9: normal plot using the median rule to detect outliers



The hypothesis and significance level retained, we do reject H_0 and conclude that the data is approximately not normally distributed at 0.05 level of significance. Since P-value = 0.029 < α =0.05

Efficiency of the MADe Method of Outlier Detection

The table of inflation rate is given after the outliers detected by the median absolute deviation method have been removed:

For 2MADe Method, We have:

Table 21

Year	Inflation rate (%)	Year	Inflation rate (%)
1981	20.9	2001	18.9
1982	7.7	2002	12.9
1983	23.2	2003	14.0
1985	5.5	2004	15.0
1986	5.4	2005	17.9
1987	10.2	2006	8.5
1990	7.5	2007	5.4
1991	13.0	2008	15.1
1996	29.3	2009	13.9
1997	8.5	2010	11.8
1998	10.0	2011	10.3
1999	6.6	2012	12.0
2000	6.9	2013	8.0

The descriptive statistics is given below:

Table 22

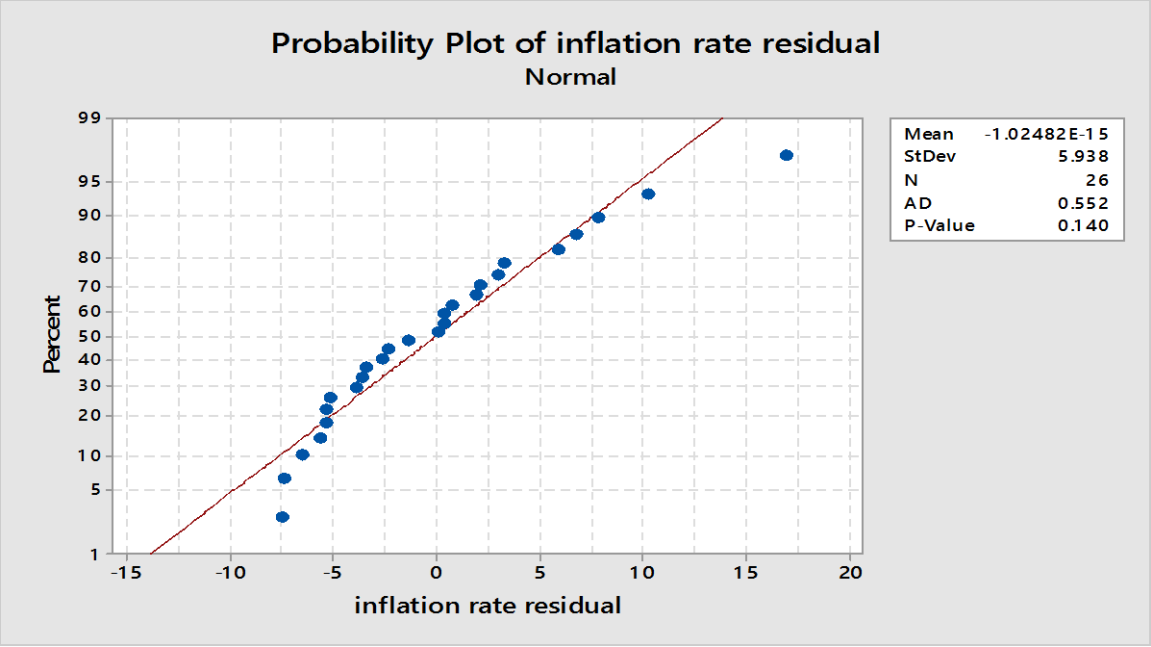
Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Inflation rate (%)	26	5.4	29.3	12.246	1.1680	5.9557
Valid N (listwise)	26					

Normality Test

The normality test for inflation rate is given below when the outlier detected by the two MADemethod has been removed.

Figure 10: normal plot for 2MADe method of outlier detection



The hypothesis and significance level retained, we do not reject H_0 and conclude that the data is approximately normally distributed at 0.05 level of significance. Since P-value = 0.140 > α =0.05

Note: The Table for 3MADe method is exactly the same with that of the 2MADe method for the given data, and same interpretation goes for it.

Interpretation

Table23:SummaryTablefor Efficiency of Different Outlier Detection Method

S/N	Detection Method	Standard Error of the Mean	Normality Result
1	2 Standard deviation	2.0732	Not Normal

2	3 Standard deviation	2.6370	Not normal
3	Z-Score Method	2.6370	Not normal
4	Modified Z-Score	1.8907	Not Normal
5	Box Plot	2.0732	Not Normal
6	Median Rule	2.0732	Not Normal
7	2 MADe	1.1680	Normal
8	3 MADe	1.1680	Normal

From the table above, we can see that the eight different methods employed in the detection of outliers in the given data perform in the same way for most of the methods.

The 3SD method and the Z-Score method can be said to be the least efficient methods of detecting outlier as they have the highest standard error of the mean and does not succeed in normalizing the data when its outliers have been removed. This can be argued to be due to the measures (mean and standard deviation) which are used in checking for outliers in this method. They were only able to detect a single outlier in the data. These methods in question are very sensitive to extreme values and the use should be discouraged.

In the same vein, we can see that the efficiency of the 2SD method, the box plot method, and the median rule are basically the same as they give the same standard error of mean of **2.0732** for the data when its outliers have been removed and even though they do not succeed in normalizing the data upon removal of the outliers at 0.05 level of significance, the data can be seen to gear towards normality. They detected same three outliers each for the three methods thus of equal efficiency

The modified Z-Score Method is the next in terms of its efficiency; it has its standard error of the mean when the outliers have been removed to be **1.8907** and even though it still does not succeed in normalizing the data at the 0.05 level of significance when outlying points are removed. It is relatively more efficient than the other five methods discussed above with respect to the given data. Four data points were detected as outliers using this method in the given data.

Both the 2MADe and 3MADe method of outlier detection can be said to be the best in the sense that it detects any single value that fails to conform at all to the general values of the given data. With the standard error of the mean when outliers have been detected being **1.1680**, it as well normalizes the data by a great deal at the 0.05 level of significance. The only problem that might be encountered in using these methods is that a good number of observations might be lost.

In general, all the methods studied can be considered a good and efficient for detecting outliers besides the 3SD and the Z-Score methods.

3. Summary and Conclusion

As shown by the work carried out in this study, each method has different ways of labeling outliers in a given data sample. The standard deviation method is very popular because of the relative ease in carrying it out thus its disadvantages are most times neglected, but it is still not a good method of testing for outliers because it uses measures like the mean and standard deviation which themselves are easily inflated in the

presence of outliers. Other methods like the median rule and the box plot make use of the median which is a robust measure and is not affected by the presence of extreme values. The most efficient methods considered which are the Modified Z-Score and the MADe rule make use of the median of the absolute deviation of the median.We know that when the data is much skewed, the box-plot method should not be used as it causes serious swamping effect. While in the presence of extreme values, the SD and Z-Score methods should not be used as they will lead to the other outliers being masked.Outlier problems are ones which have been studied quite extensively overtime. It occurs in different forms or dimensions. It can be concluded from the evidence of this study that the 3SD method and the Z-score method of outlier detection is not a good model for detecting outliers in univariate model. This can be attributed to the parameters they use for estimation of outliers in these data sets.

References

Aggarwal C.C. (2001).*Outlier Analysis*. Berlin, Kluwer Academic Publishers.

Ahmet, K. (2010). Statistical Modelling for Outlier Factors.Ozean Journal of Applied Sciences 3(1), 2010 ISSN 1943-2429

Barnett, V. and Lewis, T. (1994).*Outliers in Statistical Data*. New York, John Wiley and Sons 3rd Edition.

Ben-Gal, I. (2005), *Outlier Detection*. Glasgow, Kluwer Academic Publishers.

Breunig, M.M., Kriegel, H., Ng, R.T., and Sander, J. (2000),“ LOF: Identifying Density Based Local Outliers”. *International Conference on Management of Data, Dallas.no.4*

Carling, K. (1998), “Resistant Outlier Rules and the Non-Gaussian Case”.*Computational statistics and data analysis, vol. 33, pp 249 – 258.*

Chawla, S. and Sun, P. (2006), “Outlier Detection: Principles, Techniques and Applications.” *School of Information Technology, University of Sydney Australia.*Cousineau, D. and Chartier, S. (2010), “Outlier Detection and Treatment: A review” *International Journal of Psychological research vol.3, No.1*

Datta, P. and Kibler, D. (1995), “Learning Prototypical Concept Definition”.*In proceedings for the 12th international conferences on machine learning. Pp. 158-166, Morgan Kaufmann.*

Ferdousi, Z. and Maeda, A. (2006), “Unsupervised Outlier Detection in Time Series Data”.*Proceedings of the 22nd International Conference on Data Engineering Workshops.no. 2*

Grubbs, F.E. (1969), “Procedures for Detecting Outlying Observations in Samples”. *Technometrics 11, 1-21*

Gupta, M., Gao, J., Aggrawal, C.C., and Han, J. (2014), “Outlier Detection for Temporal Data; A Survey” *Institute of electrical and electronics engineering transactions of knowledge and data engineering vol.25, No.1* .

Hau, M.C. and Tong H., (1989), “A Practical Method for Outlier Detection in Autoregressive Time Series Modelling”. *Stochastic Hydrol. Hydraul.3, 241-260.*

Hawkins, D. (1980). *Identification of Outliers*.Massachuset, Chapman and Hall.

Heymann, S., Letapy, M., and Magnien, C. (2012), “Outskewer: Using Skewness to Spot Outliers in Samples and Time Series”. *Universite Pierre et Marie Curie, 4 Place Jussien, 75252 Paris, France.*

Hodge, V.J., and Austin, J. (2004). *A Survey of Outlier Detection Methodologies*. Sydney, Kluwer Academic Publishers.

Hsiao, C. and Tian, X. (2011),“ Intelligent Decisions: Towards Interpreting the D-algorithm”.*International Journal of Psychological research vol.3, No.2*

“Human Longevity Facts.” *http:www.myth-one.com/chapter_19.htm*

Iglewicz, B., Hoaglin, D. (1993), “How to Detect and Handle Outliers”.*ASQC Quality Press.*

Jagaddish, H.V., Koudas, N., and Muthukrishnan, S., (1999), “Mining Deviants in A Time Series Database”. *In procurement of the 25th international conference on Very Large Data Bases (VLDB), pp 102-113*

Kaya, A. (2010),“ StatisticalModelling for Outlier Factors”.*Ozean Journal of Applied Sciences vol.3(1).pp 107-121*

Kiware, S. (2010),“ Detection of Outliers on Time Series Data”. *Marquette University e-publication.*

Knorr, E.M. and Ng, R.T. (1998), “Algorithm for Mining Distance Based Outliers in Large Data Sets”.*University of British Columbia Vancouver, BC V6T I24 Canada.*

Kriegel, H.P., Kroger, P., and Zimek, A. (2010), “Outlier Detection Techniques”.*Society of Industrial and Applied Mathematics international conference on data mining.no 8*

Nare, H., Maposa D., and Lesaona, M. (2012),“ A Method for Detection and Correction of Outlier in Time Series Data.” *African Journal of Business Management.vol 3, no.2, pp 43-57*

Olewuezi, N.P .(2011), “Note on the Comparison of Some Outlier Labeling Techniques”.*Journal of Mathematics and Statistics, vol 7, pp353-355.*

Ranjit, K.P. (2001), “Some Methods of Detection of Outliers in Linear Regression Model”. *Indian Agricultural Statistics Research Institute, Library avenue, New Delhi*.

Ratcliff, R. (1993), “Methods of Dealing with Reaction Time Outliers”. *Psychological bulletin, vol. 114, pp 510 – 532.*

Rousseeuw, P. and Leroy, A. (1996), *Robust Regression and Outlier Detection*. John Wiley and Sons, 3rd Edition.

Regina, K.and Agustin, M. (2001). Seasonal Outliers in Time Series. Partially supported by the Spanish grant PB95-0299 of CICYT.

Seo, S. (2006), “A Review and Comparison of Methods for Detecting Outliers in Univariate Data sets”. *University of Pittsburgh*.

Skalak, D.B. (1994), “Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms”. *In: Machine learning proceedings of the 11th international conference. Pp 293-301.*

Tukey, J.W. (1977). *Explanatory Data Analysis*. Kyiv, Addison-Wesley.

Yamanishi, K. and Takeuchi, J. (2002), “A Unifying Framework for Detecting Outliers and Change Points From Non-Stationary Time Series Data”.*NEC Corporations, Alberta Canada.*